**Covid 19 Data Analysis**

**Business Case Scenario:** Large amount of data is available on COVID-19 for public access from government organizations and other sources like large news agencies in several countries. The data includes incidence of the pandemic such as confirmed cases, casualties for example in the US at the state level, county and at region level. Now data on vaccination allocation and distribution also is available from these sources which is continuously getting updated.

**Data Dictionary**

Center for Disease Control (CDC) a US government organization , makes available data on the number of COVID-19 cases, fatalities data-wise on each state on a daily basis.It also publishes data on allocation of vaccines from the 3 authorized pharmaceutical companies – Pfizer, Moderna and Jenssen on a weekly basis giving the number of vaccinations allocated and distributed to each state.

* *COVID-19 Cases and Deaths by State over Time:* Considered the data COVID-19 Cases and Deaths by State over Time. The data from the last week of Jan-2020 till the second week of May-2021 is taken up for the data processing and analysis exercises. The data is updated on a daily basis.
* COVID-19 Vaccine Distribution Allocations by Jurisdiction : US Food and Drug Administration has authorized 3 brands of vaccines like Pfizer, Moderna and Jenssen. Data on allocation and distribution of these vaccines is available from the last week of Decembet-2020 till second week of May-2021 and it is updated on a weekly basis.
* *Census Data :* Vaccine allocation and distribution is done based on the population data. Population data is available from the US Census organization, with the base data of 2010 and the estimated population of 2019 is taken as the basis by CDC.

**Problem statement:**

As a Data engineer, your role is to conduct a detailed data preprocessing of the above given data. You may follow the below activities to perform the analysis.

* The first step of processing the data is to clean up the records based on certain criteria.
* Next the data needs to be augmented by adding a few fields to facilitate better data analysis.
* Once the above steps are completed the specified reports need to be generated from the data.

**Approach**

* PySpark and Hive are to be used for the project.
* The data can be loaded into a PySpark SQL DataFrames creating as many DataFrames as required by the available data.
* From the census data only the required columns need to be taken and loaded into a look up table to get the population of any state based on the state code.
* A delimited file of the names of states in the US and their standard code also needs to be used as it gives a standardize way to refer to the states in all the datasets.